
Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
 Volume 25, Issues 6, pp. 5258–5264, DOI 10.12694/scpe.v25i6.3365

CHARACTER-LEVEL EMBEDDING USING FASTTEXT AND LSTM FOR BIOMEDICAL NAMED ENTITY RECOGNITION

AHMED SABAH AHMED AL-JUMAILI *AND HUDA KADHIM TAYYEH †

Abstract. Extracting biomedical entities has caught many researchers' attention in which the recent technique of word embedding is employed for such a task. Yet, the traditional word embedding architectures of Word2vec or Glove are still suffering from the 'out-of-vocabulary' (OOV) problem. This problem occurs when an unseen term might be encountered during the testing which leads to absence of embedding vector. Hence, this study aims to propose a character-level embedding through FastText architecture. In fact, handling the character-level seems a promising solution for the OOV problem. To this end, the proposed FastText architecture has been used to generate embedding vectors for the possible N-gram combinations of each word. Consequentially, these vectors have been fed to a Long Short Term Memory (LSTM) architecture for classifying the words into its biomedical classes. Using two benchmark datasets of BioCreative-II and NCBI, the proposed method was able to produce an f-measure of 0.912 and 0.918 respectively. Comparing these results with the baseline studies demonstrates the superiority of the proposed character-level embedding of FastText in terms of Biomedical Named Entity Recognition (BNER) task.

Key words: Biomedical Named Entity Recognition, FastText, Long Short Term Memory, Character-level, Out of Vocabulary.

1. Introduction. The dramatic growth of biomedical and medical data represented by publications, books, blogs and others has demonstrated the need for detecting biomedical entities. Entities like disease names, drug names, symptoms and chemical compounds are frequently occurring in biomedical resources [1], [2]. The need of recognizing these entities lies in the benefits of determining side-effects, adverse drug reactions, drug-drug interactions and other valuable information that could be mentioned implicitly or explicitly through the text. Hence, theNamed Entity Recognition task in the biomedical domain (BNER) emerged to train the machine for identifying such entities [3], [4].

The earliest research efforts on BNER were relying on engineered features such as length, position, and frequency of the term along with dictionary-based approaches [5]. However, the emergence of new sophisticated techniques such as the Word Embedding has contributed toward improving the BNER task [[6]-[9]]. Word embedding is a technique that utilizes a Neural Network architecture to predict target term given its context terms or vice versa. The main goal of such a prediction is to learn distinctive embedding vectors of the terms where such a vector would represent the term in multi-dimensional space. In this regard, terms with similar context would have similar vector representation [10]. Yet, there are various issues have been encountered by the word embedding technique such as the amount of trainable text in which word embedding requires massive text for the training in order to produce accurate vector embedding [11]. In addition, the Out-of-Vocabulary (OOV) problem was the main challenge in which an unseen term within the training might occur during the testing where the word embedding model would have no embedding for such a term [12], [13]. These challenges come from the fact that the traditional word embedding approach is dealing with word-level. Therefore, this study aims to utilize a character-level embedding using FastText architecture in order to overcome the aforementioned drawbacks.

The paper is organized as; Section 2 provides the related work, Section 3 illustrates the proposed LSTM with character mapping, Section 4 highlights the results and provide a discussion where the comparison against the baseline study is occurred, Section 5 provides the final conclusion.

*Department of Business Information Technology (BIT), College of Business Informatics, University of Information Technology and Communications, Baghdad, Iraq (asabahj@uoitc.edu.iq)

†Department of Informatics Systems Management (ISM), College of Business Informatics, University of Information Technology and Communications, Baghdad, Iraq (haljobori@uoitc.edu.iq)

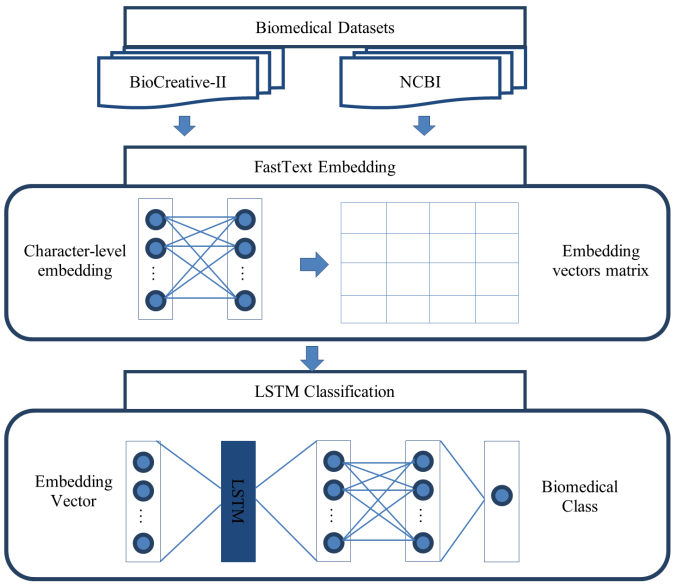

Fig. 2.1: Framework of the proposed method.

1.1. Related Work. The recent literature on BNER task has concentrated on word embedding techniques. For example, Gridach [14] have utilized a Word2Vec architecture to extract biomedical instances. Through a neural network architecture known as Long Short Term Memory (LSTM) with BioCreative-II dataset, the proposed method acquired 89.46% of f-measure. Similarly, Li and Jiang [15] have utilized the Word2Vec architecture for upon the same dataset with LSTM and got similar performance of f-measure (89.49%). Zhu et al. [16] utilized the Word2Vec with another neural network architecture known as Convolutional Neural Network (CNN) for BNER task. Based on the BioCreative-II and NCBI datasets, this study acquired an f-measure of 87.2% and 87.26% respectively. Cho and Lee [17] utilized another word embedding architecture known as Glove with LSTM for the BNER task. Using BioCreative-II and NCBI datasetsthe authors showed an f-measure of 81.44% and 85.68% respectively. Zhang and Wu [18] have proposed the Word2Vec architecture with LSTM for the BNER task. Using the BioCreative-II dataset, the proposed method obtained an f-measure of 89.94

2. Proposed Method. The proposed method of this study lies in the utilization of character-level embedding through FastText architecture. To this end, two benchmark biomedical datasets will be considered in this study including NCBI and BioCreative-II. Lastly, an LSTM architecture will process the embedding vectors generated from FastText to perform the classification of biomedical entities. Fig. 2.1 depicts these phases.

2.1. Dataset. In this study, two benchmark datasets of biomedical entities will be used. The first dataset is NCBI which introduced by Doğan et al. [19]. It concentrates on disease mentioning brought from one of the large medical sources of PubMed. The second dataset is BioCreative-II [20] consisting of genes and gene-related mentions. Table 2.1 shows the description of both datasets.

2.2. Character-level Embedding using FastText. The FastText architecture is very similar to the traditional Word2Vec in the context of examining part of text through a neural network architecture in order to predict specific target of the text. Yet, instead of handling the word-level like in Word2Vec, FastText handles the text as a character-level N-gram. In fact, handling the word-level showed various drawbacks such as the OOV problem which occurs due to the absence of vector embedding for unseen terms. Sometimes the OOV term would have derivational inflection matching within the embedding model but because the actual matching

Table 2.1: Dataset description

Dataset	Class	Description	Quantity
NCBI	UN	Disease name	8475
	O	Non-disease name	120,569
BioCreative-II	UN	Genes	15,700
	O	Non-gene	371,000

Table 2.2: One-hot encoding matrix for potential character n-gram of the 'cancers' term

	ca	an	nc	ce	er	rs
ca	1	0	0	0	0	0
an	0	1	0	0	0	0
nc	0	0	1	0	0	0
ce	0	0	0	1	0	0
er	0	0	0	0	1	0
rs	0	0	0	0	0	1

does not occur thus, no vector embedding can be brought. Assume a word embedding model that has been trained on a wide range of text contexts. Within such contexts, suppose the word 'cancer' has occurred without its derivational inflections such as 'cancers'. Handling such an inflection within the testing would lead to the OOV problem even though the model has seen similar term. Therefore, the FastText model has been emerged as a solution for this problem where the series of N-gram character of each word will be trained.
FastText architecture proposed by Facebook to process sequences of N-gram characters for every individual word and averaging the resulted vectors into a single embedding vector [21]. To illustrate the way of doing such an embedding, assume a term of 'cancers', FastText will produce a sparce matrix known as one-hot encoding matrix which is typical to the way Word2Vec works. Yet, rather than focusing on word-level contexts, FastText addresses the N-gram character sequence of the word. Table 2.2 shows an example using 'cancers' term.

Hence, the FastText architecture will process the potential character N-gram in order to predict specific N-gram characters. Like the Word2Vec, FastText will train the model and tune the weights in order to get matching between the prediction and the actual values. In this regard, the hidden neurons will articulate the embedding vector for the targeted N-gram characters. Fig. 2.2 represents a simple architecture of FastText where the N-gram character sequences of the word 'cancers' including 'ca', 'an', 'nc', 'ce', and 'er' are being processed in the input to predict the last N-gram sequence of 'rs'.

2.3. LSTM. Once the FastText model is being built and trained on the two corpora of NCBI and BioCreative, an LSTM architecture will be used to classify the words into its biomedical class label. LSTM is an architecture that has been built upon the Recurrent Neural Network (RNN) architecture which introduced the recurrent feedback connections [22]. Additionally, LSTM has extra components of memory and forget gate in which the contextual information is being saved and insignificant information is being forgotten [[23]-[25]]. LSTM has been widely used for sequence and time series data classification and prediction. Fig. 2.3 depicts the architecture of LSTM used in this study.

As shown in Fig. 2.3, the proposed LSTM will process the embedding vector of each word which has been generated by FastText. Apparently, the dimension of such a vector is 100. After that, an LSTM layer with a dimension of 32 will be utilized with a dropout in order to prevent overfitting. Consequentially, a dense layer or so-called a fully connected layer with a dimension of 64 will be utilized. Lastly, the output layer will be supplemented with a Softmax in order to articulate the biomedical class label of the word.

3. Results and discussion. Prior to show the experimental results, it is necessary to consider the experiment settings of both the FastText model and the LSTM architecture. Following subsections show the experiment settings and experiment results.

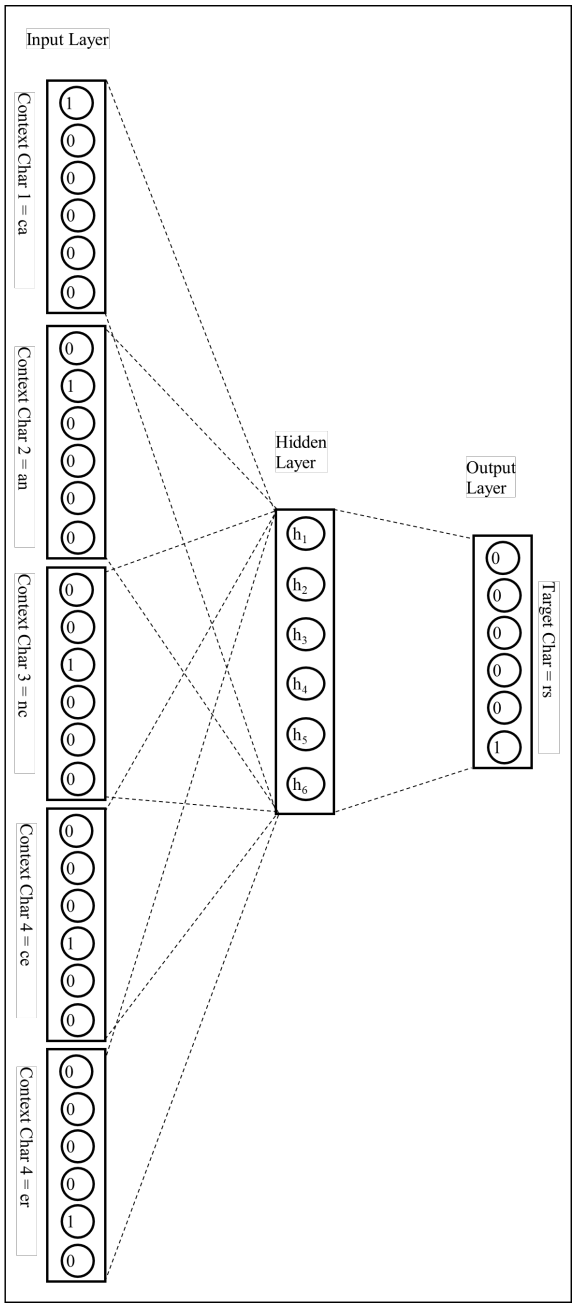

Fig. 2.2: Simple architecture of FastText.

3.1. Experiment Setting. Table 3.1 shows the hyperparameters of FastText model. Whereas Table 3.2 shows the hyperparameters of LSTM architecture.

3.2. Experiment Results. The results will be examined based on the three metrics of precision, recall and f-measure. Table 3.3 shows the results of applying FastText and LSTM for both datasets.

As shown in Table 3.3, the proposed method had the ability to identify BNEs within the NCBI dataset with a precision of 0.923, a recall of 0.901, and an f-measure of 0.911. Similarly, for the BioCreative-II dataset,

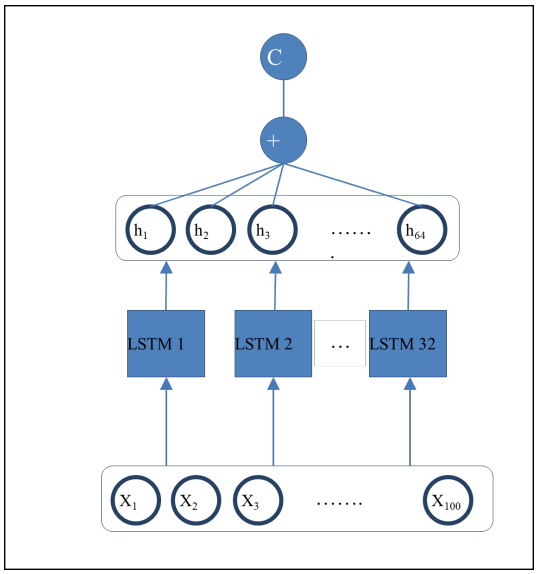

Fig. 2.3: LSTM architecture.

Table 3.1: Hyperparameter of FastText model

Hyperparameter	Description
Dimension	100
Window size	5
Number of epochs	10
Architecture	Skip-gram

Table 3.2: Hyperparameter of LSTM architecture

Hyperparameter	Description
Batch size	64
LSTM layer	32
Dropout	0.25
Dense layer	64
Number of epochs	100
Optimizer	Adam

Table 3.3: Results of BNER using FastText and LSTM

Dataset	Precision	Recall	F-measure
NCBI	0.92301	0.90105	0.91189
BioCreative-II	0.91929	0.90677	0.91298

the proposed method achieved 0.919, 0.906 and 0.912 for precision, recall and f-measure respectively.

3.3. Discussion. As depicted earlier, the proposed method has outperformed both studies of Gridach [14] and Li and Jiang [15] who used LSTM with Word2Vec upon the BioCreative-II dataset and obtained an f-measure of 0.894. On the other hand, the proposed method has superior performance of f-measure compared to

the study of Zhu et al. [16] who used CNN with Word2Vec on BioCreative-II and NCBI datasets and obtained 0.872 of f-measure respectively. Additionally, the study of Cho and Lee [17] who used LSTM and Glove for both BioCreative-II and NCBI datasets and achieved an f-measure of 0.814 and 0.856 respectively is still having lower performance compared to the proposed method. Lastly, the study of Zhang and Wu [18] who used LSTM with Word2Vec for BioCreative-II dataset and obtained 0.899 is still having lower f-measure compared to the proposed method. Generally, the proposed FastText embedding has contributed toward enhancing the recognition of BNEs. This is due to the character-level treatment that has reduce the OOV problem.

4. Conclusion. This paper has presented a character-level embedding approach using FastText. The embedding has utilized the N-gram character permutations in order to give a distinctive embedding vector for each combination. These vectors have processed through LSTM to classify the words into its biomedical classes. Experimental results demonstrated for the proposed method over the state-of-the-artin terms of f-measure using two well-known datasets. For future directions, utilizing a pretrained FastText embedding model might contribute toward enhancing the classification performance.

Acknowledgements. This study has been supported by the University of Information Technology and Communications.

REFERENCES

[1] V. Kocaman, and D. Talby, *Biomedical named entity recognition at scale* In: Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part I. Springer International Publishing, 2021. p. 635-646.

[2] J. Ravikumar, and K. P. Ramakanth, *Machine learning model for clinical named entity recognition*, International Journal of Electrical and Computer Engineering, vol. 11, no. 2, pp. 1689, 2021.

[3] Z. Chai, H. Jin, S. Shi, S. Zhan, L. Zhuo, and Y. Yang, *Hierarchical shared transfer learning for biomedical named entity recognition*, BMC bioinformatics, vol. 23, no. 1, pp. 1-14, 2022.

[4] K. Mrhar, and M. Abik, *Towards optimize-ESA for text semantic similarity: A case study of biomedical text*, International Journal of Electrical and Computer Engineering, vol. 10, no. 3, pp. 29-34, 2020.

[5] B. Song, F. Li, Y. Liu, and X. Zeng, *Deep learning methods for biomedical named entity recognition: a survey and qualitative comparison*, Briefings in Bioinformatics, vol. 22, no. 6, pp. bbab282, 2021.

[6] Z. Nasar, S. W. Jaffry, and M. K. Malik, *Named entity recognition and relation extraction: State-of-the-art*, ACM Computing Surveys (CSUR), vol. 54, no. 1, pp. 1-39, 2021.

[7] S. S. Lwin, and K. T. Nwet, *Myanmar news summarization using different word representations*, International Journal of Electrical & Computer Engineering (2088-8708), vol. 11, no. 3, 2021.

[8] A. Hadioui, Y. B. Touimi, N.-e. E. Faddouli, and S. Bennani, *Intelligent machine for ontological representation of massive pedagogical knowledge based on neural networks*, International Journal of Electrical & Computer Engineering (2088-8708), vol. 11, no. 2, 2021.

[9] M. A. Fauzi, *Word2Vec model for sentiment analysis of product reviews in Indonesian language*, International Journal of Electrical and Computer Engineering, vol. 9, no. 1, pp. 525, 2019.

[10] Y. Goldberg, and O. Levy, *word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method*, arXiv preprint arXiv:1402.3722, 2014.

[11] Y. Chen, C. Zhou, T. Li, H. Wu, X. Zhao, K. Ye, and J. Liao, *Named entity recognition from Chinese adverse drug event reports with lexical feature based BiLSTM-CRF and tri-training*, Journal of biomedical informatics, vol. 96, pp. 103252, 2019.

[12] M. Khalifa, and K. Shaalan, *Character convolutions for Arabic Named Entity Recognition with Long Short-Term Memory Networks*, Computer Speech & Language, vol. 58, pp. 335-346, 2019/11/01/, 2019.

[13] R. E. Ramos-Vargas, I. Román-Godínez, and S. Torres-Ramos, *Comparing general and specialized word embeddings for biomedical named entity recognition*, PeerJ Computer Science, vol. 7, pp. e384, 2021.

[14] M. Gridach, *Character-level neural network for biomedical named entity recognition*, Journal of Biomedical Informatics, vol. 70, pp. 85-91, 2017.

[15] L. Li, and Y. Jiang, *Biomedical named entity recognition based on the two channels and sentence-level reading control conditioned LSTM-CRF*, In: 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, 2017. p. 380-385.

[16] Q. Zhu, X. Li, A. Conesa, and C. Pereira, *GRAM-CNN: a deep learning approach with local context for named entity recognition in biomedical text*, Bioinformatics, vol. 34, no. 9, pp. 1547-1554, 2017.

[17] H. Cho, and H. Lee, *Biomedical named entity recognition using deep neural networks with contextual information*, BMC Bioinformatics, vol. 20, no. 1, pp. 735, 2019.

[18] L. Zhang, and H. Wu, *Medical Text Entity Recognition Based on Deep Learning*, In: Journal of Physics: Conference Series. IOP Publishing, 2021. p. 042209

[19] R. I. Doğan, R. Leaman, and Z. Lu, *NCBI disease corpus: a resource for disease name recognition and concept normalization*, Journal of biomedical informatics, vol. 47, pp. 1-10, 2014.
[20] L. Smith, L. K. Tanabe, R. J. Ando, C.-J. Kuo, I.-F. Chung, C.-N. Hsu, Y.-S. Lin, R. Klinger, C. M. Friedrich, and K. Ganchev, *Overview of BioCreative II gene mention recognition*, Genome biology, vol. 9, no. Suppl 2, pp. S2, 2008.
[21] H. Pylieva, A. Chernodub, N. Grabar, and T. Hamon, *Improving automatic categorization of technical vs. Laymen medical words using FastText word embeddings*, In 1st International Workshop on Informatics & Data-Driven Medicine (IDDM 2018)
[22] K. A. Wahdan, S. Hantoobi, S. A. Salloum, and K. Shaalan, *A systematic review of text classification research based ondeep learning models in Arabic language*, Int. J. Electr. Comput. Eng, vol. 10, no. 6, pp. 6629-6643, 2020.
[23] Z. Ferdoush, B. N. Mahmud, A. Chakrabarty, and J. Uddin, *A short-term hybrid forecasting model for time series electrical-load data using random forest and bidirectional long short-term memory*, International Journal of Electrical & Computer Engineering (2088-8708), vol. 11, no. 1, 2021.
[24] A. Nasser, and H. Al-Khazraji, *A hybrid of convolutional neural network and long short-term memory network approach to predictive maintenance*, International Journal of Electrical & Computer Engineering (2088-8708), vol. 12, no. 1, 2022.
[25] S. Krishnan, P. Magalingam, and R. Ibrahim, *Hybrid deep learning model using recurrent neural network and gated recurrent unit for heart disease prediction*, International Journal of Electrical & Computer Engineering (2088-8708), vol. 11, no. 6, 2021.

Edited by: Mustafa M Matalgah
Special issue on: Synergies of Neural Networks, Neurorobotics, and Brain-Computer Interface Technology: Advancements and Applications
Received: Feb 13, 2024
Accepted: Jul 18, 2024